Automated Extraction and Normalization of Findings from Cancer-Related Free-Text Radiology Reports

Burke W. Mamlin, M.D.[†], Daniel T. Heinze, Ph.D.[‡], Clement J. McDonald, M.D.[†]
[†]Regenstrief Institute for Health Care, Indianapolis, Indiana
[‡]A-Life Medical, Inc., San Diego, California

ABSTRACT

We describe the performance of a particular natural language processing system that uses knowledge vectors to extract findings from radiology reports. LifeCode® (A-Life Medical, Inc.) has been successfully coding reports for billing purposes for several years. In this study, we describe the use of LifeCode® to code all findings within a set of 500 cancer-related radiology reports against a test set in which all findings were manually tagged. The system was trained with 1400 reports prior to running the test set. Results: LifeCode® had a recall of 84.5% and precision of 95.7% in the coding of cancer-related radiology report findings. Conclusion: Despite the use of a modest sized training set and minimal training iterations, when applied to cancer-related reports the system achieved recall and precision measures comparable to other reputable natural language processors in this domain.

INTRODUCTION

A wealth of clinical data exists within dictated clinical notes and other electronic medical text.[1] Given the terabytes of electronic narrative being produced annually, the need for tools to extract coded data from these reports is clear. The time-resource demands for human coding are significant and human coders can fall behind the rate at which narratives are produced. Natural language processing (NLP) tools have been studied and used to address this problem for several years;[2] in fact, commercial systems such as A-Life's LifeCode® are already using NLP to extract billing codes from dictated reports. However, the potential uses for NLP go far beyond billing. Data within narrative reports can be used for data mining[3], research queries[4], patient management[5], computer-generated reminders[6], guidelines[7], detecting comorbidities[8], detecting adverse events[9], quality assessment[10] and decision support.[11]

In this study, we evaluate the extension of a commercially available product from billing code discovery to complete encoding of narrative reports; specifically, cancer-related chest x-ray reports. We describe the evaluation of LifeCode® according to the recommended methods for evaluating NLP in the clinical domain.[12]

BACKGROUND

Natural language processing technology has been around for many years.[13] Although significant progress has been made, there is still room for improvement. For example, most implemented medical NLP systems quote recall in the 80-85% range and precision of 95-99%.[14] Though this level of performance is not perfect, it may be reasonable for real-world application, given that human coders fall within the same range when compared to one another.[15-17]

Speaking broadly, current approaches to medical NLP can be classified as predominantly statistical or predominantly symbolic. Statistical systems operate on the basis of word proximity and frequency and partition the decision space using techniques that include naïve Bayes, support vectors, n-grams and neural networks. Systems that could be classified as primarily statistical include CodeRyte™, MEDSYNDIKATE, and PlatoCode™. Symbolic systems treat words as symbols within a grammar that defines the allowable associations between concepts (as opposed to proximity). Furthermore, symbolic systems usually include some form of a knowledge base for validation and classification of symbolic phrases. Systems that can be classified as primarily symbolic are MedLEE and LifeCode®.[18] Although there are statistical techniques for building symbolic grammars for natural languages, we are not aware of any medical NLP systems that are so constructed. Spyns provides an overview of some NLP techniques for medical applications.[19]

LifeCode® combines NLP and a medical coding expert system in a commercial product that extracts and normalizes demographic and clinical information from free-text clinical records. LifeCode® was initially applied to the extraction of billing codes from Emergency Medicine reports. Since October 2000, it has been performing this same task for radiology reports. The program is implemented largely in C++, combining dozens of specialized components. A detailed description of LifeCode® is provided elsewhere.[20] In overview, LifeCode® uses an array of about three dozen specialized parsers to perform document segmentation and various types of phrase normalization. Several chunking grammars are used for sentence parsing. These grammars differ according to the type of information extraction, e.g. findings and diagnoses versus procedures. LifeCode's knowledge bases are of three distinct types: knowledge vectors, rule bases, and semi-knowledge. "Knowledge vectors" are vector space knowledge bases with a representation similar to feature vectors and are used to map symbolic

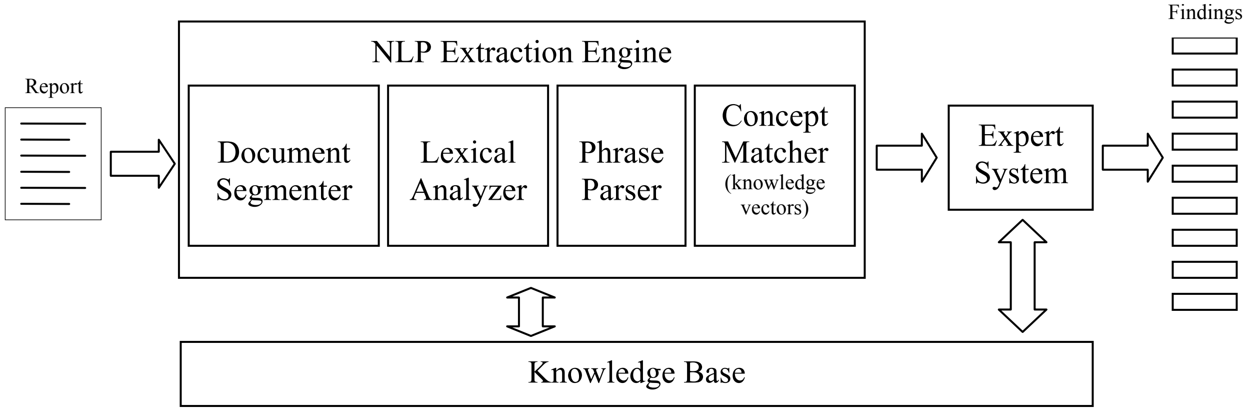

Figure 1. LifeCode® Architecture. The NLP Extraction Engine breaks reports up into concept-level phrases and matches them to known concepts using knowledge vectors.

phrases to standard or proprietary codes. Rule bases are used to remove ambiguous and redundant codes, combine codes and apply coding logic. "Semi-knowledge" is a representation that allows for a gradual degradation of performance along the boundaries of the system's knowledge. At the time of writing this paper, the rule bases and semi-knowledge have not yet been included in the project of record, although they are used in the commercial LifeCode® applications.

When dealing with a single sentence, the LifeCode® core engine references the linguistic and medical knowledge base, on average, 50,000 times (ranging from several thousand to several million times). A large table storing partial results during the vector analysis allows these calculations to be performed in a reasonable amount of time.

METHODS

We used chest x-ray reports to train and then test the performance of the LifeCode® system. This study received exempt status from our Institutional Review Board. All chest x-ray reports used in this study were previously scrubbed using a de-identification tool.[21] Ages and gender were replaced with random values and all other identifying information was replaced with placeholders (e.g., "*Spoke with Dr. NAME from INSTITUTION at TIME on DATE.*"). Randomly assigned, unique identifiers were used to label the reports.

Reports. All chest x-ray reports (n=26,778) generated during 2002 at a county hospital, Wishard Memorial Hospital in Indianapolis, Indiana, were considered for this study. Because this effort was part of a National Cancer Institute-funded project to identify clinical findings that might distinguish cancer phenotypes, we focused on the detection of cancer-related findings such as *metastases*. In order to favor these cases and exclude normal studies, we filtered the 26,778 reports by searching for the presence of the following words: *adenocarcinoma, cancer, Hodgkin's, leukemia, malignancy, malignant, mass, mesothelioma, metastases, metastasis, metastatic, neoplasm, neoplastic, nodule, Non-Hodgkin's, pneumonectomy, suspicious,* and *tumor.* Since these words could be mentioned in a negative context within normal chest x-rays (e.g., "*Mediastinal structures show no evidence of mass.*"), we further filtered the reports to those *not* containing the following phrases: *clear chest, no active, no acute, no cardiopulmonary, normal chest, normal exam,* and *unremarkable*. After filtering the chest x-rays, there were 3,015 chest x-ray reports remaining.

These 3,015 chest x-ray reports were dictated by 26 radiologists working at Wishard Memorial Hospital Radiology Department, using PowerScribe® (Dictaphone™, Stratford, Connecticut) to perform immediate speech recognition prior to editing and signing the report. The reports were drawn from the Regenstrief Medical Record System[22] as simple text with minimal structure. Two sections were readily separated: the *narrative* and the *impression*. The *narrative* included findings (i.e., reason for the study), comparison, and detailed description of the exam. The *impression* contained a summary of findings dictated by the radiologist. Headings within these two sections (e.g., "*Findings:*", "*Comparison:*") were often, but not always, provided as part of the dictation.

Training. Four training sets of scrubbed chest x-ray reports – a total of 1400 reports – were sent to A-life for training. In each case, the reports were randomly selected from the pool of 3,015 chest x-ray reports. The first three training sets consisted of 100 to 150 reports each; the last training set contained 1000 reports. The first two training sets were subject to two training iterations; the third and fourth training sets were, due to time constraints, subjected to only one iteration (i.e. system updates made on the basis of the first iteration were not subsequently tested and validated). The system was trained by processing a set of reports, manually reviewing the results and manually making knowledge base corrections as indicated.

Testing. Five hundred randomly selected chest x-ray reports were manually coded by a board-certified internist (BWM), who is not himself a developer of LifeCode®. These reports were drawn from the pool of 3,015 filtered reports and were not part of any training set. They had not been seen by A-Life. Before submitting the reports to A-Life, BWM read and manually tagged all findings within the reports. Each tag included a finding type (*negative, normal, positive, possible, probably, history, mild, stable, worse, improved, appliance,* or *recommendation*) and any modifying words from within the report. For example, in the sentence "*Mediastinal structures show a mildly tortuous aorta, but no definite mass*." two findings were tagged: *aorta (type: mild, modifiers: tortuous)* and *mass (type: negative, modifiers: none)*. Findings were tagged using a simple browser-based annotation tool, consequently findings and modifiers were all based on words within the narrative. After this process, the tags were all removed and this test set of scrubbed reports was delivered to A-Life in California. We had pre-agreed that the investigators would not use any of the information in this test set to adjust the code and the coded reports would be returned immediately following processing.

Evaluation. BWM manually compared the machine-based coding with the reference set. For each report, human- and computer-generated codes were listed side-by-side with the complete narrative displayed above them. All matching codes were linked. Codes incorrectly generated by the computer were tagged as false positives. Codes missed by the computer were tagged as false negatives. Because the A-life system does not yet code recommendations, the recommendations that were manually tagged were not included in the analysis.

RESULTS

We manually coded 500 reports and compared the manual codes to the codes generated by the LifeCode® system. It took BWM approximately 20 hours to manually code the reports (averaging slightly less than 2.5 minutes per report). A-Life returned the reports within six minutes and reported that LifeCode® finished coding the same reports in less than three minutes (average 0.34 seconds per report).[*] The 500 reports used in the reference set contained 4,901 sentences. There were 130 (26%) "normal" reports (no positive findings within the impression). The distribution of the most common findings is reported in Table 1.

There were 5,263 manually coded entries within the reports (8,021 modifiers); 254 of the coded entries were recommendations or duplicate references to the same finding, leaving 5,009 non-redundant potential findings. In 130 cases, findings were mentioned within a "rule-out" statement in the reason for the study; these were not manually coded, but were recognized by the computer as "possible" findings. Therefore, there were 5,139 total codes considered to be "correct" findings (i.e. the closest to a "gold standard") for this study. On average, there were ten codes per report (seven in the narrative and three in the impression).

Out of 5,139 possible findings, 4,347 were coded correctly and 792 were missed; 195 incorrect codes were generated. The computer's recall was 84.6% and precision was 95.7% .

Table 1. Distribution of the top 20 findings[*] (within reason for study, narrative, or impression) for chest x-ray reports (*n=500*).

Opacity	29.6%
Cough	21.0%
Cardiomegaly	15.0%
Airspace disease	14.8%
Pulmonary/pleural node	14.8%
Pulmonary atelectasis	13.8%
Chest pain, unspecified	13.4%
Pleural effusion	13.4%
Mass	13.0%
Heart dilated	12.6%
Node	12.2%
Other dyspnea and respiratory abnormalities	11.6%
Hilar, prominence	8.6%
Shortness of breath	8.4%
Pulmonary/pleural cancer	8.2%
Appliance: tube, endotracheal	7.8%
Low lung volume	7.8%
Granuloma, calcification	7.6%
Pulmonary/pleural infiltrate	7.6%
Swelling, mass, or lump in chest	7.6%

[*]Radiology reports were pre-selected to favor cancer-related findings

Subanalyses were performed to evaluate performance within the two sections of our reports and among normal versus abnormal findings. Within the *narrative* section, the computer's recall was 86.6% and precision was 95.9%. When only *impressions* were considered, the computer's recall and precision were 76.8% and 95%, respectively. Considering only normal findings (*normal, stable,* or *negative* modifiers), LifeCode® recall and precision were 83.6% and 98.6%, respectively. Among positive findings (*positive* or *worse* modifiers), the recall and precision were 85.6% and 82.6%, respectively.

[*] Additional time was taken decompressing the files, recompressing the files, and sending them via e-mail.

DISCUSSION

Despite the use of a modest-sized training set (only 1400 reports) and a limited number of training iterations, the observed performance of LifeCode® for cancer-related x-ray reports is comparable to other medical NLP systems and to human coders. We found recall 84.6% and precision 95.7% overall; however, during evaluation of the system, we noticed a small set of errors that occurred throughout the documents. These errors resulted from technical details (e.g. weighting errors). There were four situations in particular that can easily be remedied and were responsible for a significant portion of the errors. (1) The phrase "*No evidence of acute disease*" at the end of several reports was consistently miscoded; (2) terse impressions comprised solely of the word "*normal*" or "*clear*" were not coded by the computer; (3) the phrase "*Heart size and vessels were normal*" was common and only the heart size was coded by the computer (missing normal vessels); and, for technical reasons, (4) a single term ("crackles") was erroneously added to a quarter of the reports. Each of these errors is easily remedied and represents a system rather than report-specific problem. If we adjusted for these four errors, the recall would rise somewhat from 84.6% to 87.4% and the precision from 95.7% to 98.6%. Not surprisingly, the benefit of these fixes affected normal findings (recall/precision from 83.6%/98.6% to 89.1%/99.5%) while only affecting the precision for abnormal findings (85.6%/82.6% to 85.6%/94.1%).

Several reasons can be offered for the drop in recall in the *impression* section compared to other parts of the report. First, because the emphasis to date has been on developing a coding system for findings that are not covered by other standard coding systems, only a few of the more common cancer codes are currently entered in the LifeCode® knowledge bases. Because the *impression* section favors diagnoses over findings, we would expect that, in a test set oriented toward cancer patients, sensitivity would be lower for the impressions. A-Life plans to enter the more complete set of oncology codes after the findings knowledge bases become more stable. Second, as noted earlier, the single words *"normal"* and *"clear"*, when standing alone, do not have a code within the system. For those 26% of the reports that were normal studies, *"normal"* and *"clear"* are not unusual as the impression. Finally, given the generally terse nature of impression sections, grammar and punctuation errors tend to have more significant adverse effects than in the wordier findings sections. The semi-knowledge capability, which will eventually be included as a part of this project, is, in part, directed at addressing this type of problem.

A-Life developed the nomenclature used by LifeCode® in this project largely on the basis of reviewing reports and researching the literature on radiology. More than 1,500 core findings codes have been developed for characterizing chest studies and the number continues to grow. ICD-9-CM codes are also used, but hierarchically they are subordinated to the new findings codes which are more specific. Within the findings codes, a hierarchy of specificity also exists, but it is not currently exercised because the over generation of codes is useful for analysis during training. Beyond the core findings codes and ICD-9-CM codes, there are attributes that can be assigned to each code. These include a history attribute (*current*, *past*, *family*), certainty (*not evident, possible, probable, alleged, denied, inferred, other person*), change (*new, worse, stable, improved, gone*), severity (*mild, moderate, severe*) and normalcy (*normal, borderline, abnormal*). Several other useful attributes that are yet to be added include *increased/decreased*, *status post*, etc. A-Life is in the process of making the anatomical location an attribute code that will appear if the location of a finding can be identified with specificity greater than that represented in the finding code.

This study has several limitations. First of all, we consider only chest x-ray reports and the reports used were skewed toward cancer-related diagnoses. While this could bias our findings, cancer-related chest x-rays are typically complex and, if anything, should make our evaluation overly conservative.

Only one physician was used to mark up the reference set; the same physician performed the linking of computer and human codes. When 20 out of the 500 test reports were randomly re-coded manually two weeks later, test-retest reliability was greater than 98%. This demonstrates a high intra-rater reliability; however, inter-rater rater reliability (e.g., comparison of at least three expert raters) was not done and would likely reduce the reliability of our results.

We cannot make conclusions about the generalizability of these findings. While we plan on testing reports from other institutions, applying LifeCode® to reports from another institution will likely involve revision of the NLP engine to accommodate the local differences in the vernacular.[23]

The dictations were produced using a speech recognition engine. This introduced occasional errors and sentence fragments within the reports that could have interfered with our results. In the worst cases, the radiologist dictated "*Heart and hilar vessels are normal*" and the speech recognition engine transcribed "*Heart and hilar mass are normal.*" This led the NLP engine to code a "normal" hilar mass. While improvements in the NLP engine will help detect and avoid these problems, we feel that speech recognition engines are likely to only increase in usage among radiology departments and, therefore, only help to demonstrate how the system can perform in real-life situations.

CONCLUSION

Despite minimal training from our reports, when applied to cancer-related reports, LifeCode® performed at a level comparable to existing NLP systems. We are encouraged that the system will only improve with further adjustments and additional training sets. LifeCode® represents a viable alternative for the extraction of medical findings from cancer-related narrative chest x-ray reports. Further study is needed to confirm these findings among a more generalized set of reports.

As a result of this work, we now have a reasonable corpus of 500 radiology reports with tagged findings that can be re-used for evaluation and improvements of other medical NLP tools. This corpus is available upon request (bmamlin@regenstrief.org).

ACKNOWLEDGEMENTS

This research was performed at the Regenstrief Institute for Health Care in Indianapolis, Indiana and was funded, in part, by National Cancer Institute grant U01 CA91343, a Cooperative Agreement for The Shared Pathology Informatics Network, and by National Library of Medicine grant T15 LM07117-06.

REFERENCES

1. Tierney WM, Overhage JM, McDonald CJ. Toward electronic medical records that improve care. Annals of Internal Medicine. 1995;122(9):725-6.
2. Haug PJ, Ranum DL, Frederick PR. Computerized extraction of coded findings from free-text radiologic reports. Work in progress. Radiology. 1990;174(2):543-8.
3. Heinze DT, Morsch ML, Holbrook J. Mining Free-Text Medical Reports. Proceedings of the AMIA Annual Symposium 2002:254-258.
4. Libbus B, Rindflesch TC. NLP-Based Information Extraction for Managing the Molecular Biology Literature. Proceedings of the AMIA Annual Symposium 2002:449-445.
5. Gundersen ML, Haug PJ, Pryor TA, van Bree R, Koehler S, Bauer K, et al. Development and evaluation of a computerized admission diagnoses encoding system. Comput Biomed Res 1996;29(5):351-72.
6. Zingmond D, Lenert LA. Monitoring free-text data using medical language processing. Computers & Biomedical Research. 1993;26(5):467-81.
7. Lenert LA, Tovar M. Automated linkage of free-text descriptions of patients with a practice guideline. Proceedings - the Annual Symposium on Computer Applications in Medical Care. 1993:274-8.
8. Chuang J-H, Friedman C, Hripcsak G. A Comparison of the Charlson Comorbidities Derived from Medical Language Processing and Administrative Data. Proceedings of the AMIA Annual Symposium 2002:160-164.
9. Bates DW, Evans RS, Murff H, Stetson PD, Pizziferri L, Hripcsak G. Detecting Adverse Events Using Information Technology. J Am Med Inform Assoc 2003;10(2):115-128.
10. Lyman M, Sager N, Tick L, Nhan N, Borst F, Scherrer JR. The application of natural-language processing to healthcare quality assessment. Med Decis Making 1991;11(4 Suppl):S65-8.
11. Knirsch CA, Jain NL, Pablos-Mendez A, Friedman C, Hripcsak G. Respiratory isolation of tuberculosis patients using clinical guidelines and an automated clinical decision support system. Infect Control Hosp Epidemiol 1998;19(2):94-100.
12. Friedman C, Hripcsak G. Evaluating natural language processors in the clinical domain. Methods of Information in Medicine. 1998;37(4-5):334-44.
13. Hripcsak G, Friedman C, Alderson PO, DuMouchel W, Johnson SB, Clayton PD. Unlocking clinical data from narrative reports: a study of natural language processing. Ann Intern Med 1995;122(9):681-8.
14. Hripcsak G, Austin JH, Alderson PO, Friedman C. Use of natural language processing to translate clinical information from a database of 889,921 chest radiographic reports. Radiology. 2002;224(1):157-63.
15. Hripcsak G, Kuperman GJ, Friedman C, Heitjan DF. A reliability study for evaluating information extraction from radiology reports. Journal of the American Medical Informatics Association. 1999;6(2):143-50.
16. Morris WC, Heinze DT, Warner Jr. HR, Primack A, Morsch AEW, Sheffer RE, et al. Assessing the Accuracy of an Automated Coding System in Emergency Medicine. Proceedings - the Annual Symposium on Computer Applications in Medical Care. 2000.
17. Friedman C, Hripcsak G, Shablinsky I. An evaluation of natural language processing methodologies. Proceedings/AMIA Annual Symposium. 1998:855-9.
18. Friedman C. A broad-coverage natural language processing system. Proceedings/AMIA Annual Symposium. 2000:270-4.
19. Spyns P. Natural language processing in medicine: an overview. Methods of Information in Medicine. 1996;35(4-5):285-301.
20. Heinze DT, Morsch ML, Sheffer RE, Jimmink MA, Jennings MA, Morris WC, et al. LifeCode: A Deployed Application for Automated Medical Coding. AI Magazine 2001;22(2):76-88.
21. Thomas SM, Mamlin B, Schadow G, McDonald CJ. A Successful Technique for Removing Names in Pathology Reports Using an Augmented Search and Replace Method. Proceedings of the AMIA Annual Symposium 2002:777-781.
22. McDonald CJ, Overhage JM, Tierney WM, Dexter PR, Martin DK, Suico JG, et al. The Regenstrief Medical Record System: a quarter century experience. Int J Med Inf 1999;54(3):225-53.
23. Hripcsak G, Kuperman GJ, Friedman C. Extracting findings from narrative reports: software transferability and sources of physician disagreement. Methods of Information in Medicine. 1998;37(1):1-7.